Hybrid sampling and Random Forest based Machine Learning Approach for Software Defect Prediction

Md. Anwar Hossen[1], Md. Shariful Islam[1], Nurhafizah Abu Talip Yusof[2], Md. Sakib Rahman[1], Fatema Siddika[3], Mostafijur Rahman[1], Sabira Khatun[2], Mohamad Shaiful Abdul Karim[2], S M Hasan Mahmud[1]

[1] Department of Software Engineering, Daffodil International University, Bangladesh.
[2] Faculty of Electrical and Electronics Engineering, University Malaysia Pahang, Malaysia.
[3] Department of Computer Science and Engineering, Jagannath University, Bangladesh.
Email: anwar.swe@diu.edu.bd

Abstract.
The software has turn into an imperious part of human's life. In the recent computing era, many large-scale complex network systems and millions of modern technological devices produce a huge amount of data every second. Among these data, the amount of imbalanced data is relatively excessive. The machine learning model is miss leaded by these imbalanced data. Software Defect Prediction (SDP) is a standout amongst the most helping exercises during the testing phase. The estimated cost of finding and fixing defects is approximately billions of pounds per year. To reduce this problem, software defect prediction has come forth but need fine tuning to have expected efficiency. In this chapter, we have proposed a new model based on machine learning approach to predict software defect and identify the key factors that may help the software engineer to identify the most defect-prone part of the system. The proposed model works as follows. First, need to remove highly correlated features and turn all the feature in the same scale using the scaling feature approach. Second, we have used Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic (ADASYN) and Hybrid sampling method to balance highly imbalanced datasets. Third, Random Forest Importance and Chi-square algorithms are chosen to find out the factors which have high effect on software defect. Cross validation is used to remove overriding problem. Scikit-learn library is used for machine learning algorithms. Pandas library is used for data processing. Matplotlib, and PyPlot are used for graph and data visualization respectively. The hybrid sampling method and Random Forest (RF) algorithms achieved the highest prediction accuracy about 93.26% by showing its superiority.

Keywords: Software Defect Prediction, Machine Learning, Imbalanced Dataset, Chi Square, Random Forest importance.

Acknowledgement

This research work is supported by research grant RDU1703236 funded by Universiti Malaysia Pahang, http://www.ump.edu.my/. The authors would also like to thank the Faculty of Electrical & Electronics Engineering, Universiti Malaysia Pahang for financial support.
4626-4636 (2009).